Method guidelines for systematic reviews in the Cochrane Collaboration back review group for spinal disorders

van Tulder, M.W.; Assendelft, W.J.J.; Koes, B.W.; Bouter, L.M.

published in
Spine
1997

DOI (link to publisher)
10.1097/00007632-199710150-00001

document version
Publisher's PDF, also known as Version of record

SPINE Volume 22, Number 20, pp 2323–2330

Method Guidelines for Systematic Reviews in the *Cochrane Collaboration* Back Review Group for Spinal Disorders

Maurits W. van Tulder, PhD,* Willem J. J. Assendelft, MD, PhD,* Bart W. Koes, PhD,* Lex M. Bouter, PhD,* and the Editorial Board of the Cochrane Collaboration Back Review Group†

Review articles offer clinicians and health policy makers the opportunity to cope with the exponentially increasing number of medical publications. For a long time, clinicians and others have been provided with evidence from unsystematic narrative reviews, but the current interest in evidence-based medicine has led to an extensive increase in the publication of systematic reviews. Because the randomized controlled trial (RCT) is generally considered to be the paradigm of intervention research, which implies that the strongest scientific proof of the effectiveness of an intervention is provided by RCTs, most systematic reviews are confined to evaluating RCTs. Reports of RCTs should be accurate and complete so that readers can evaluate the internal and external validity of the trial. Recently, several leading medical journals (BMJ, JAMA, Lancet, N Engl J Med) adopted recommendations for the reporting of RCTs (the CONSORT statement) by an expert panel that consisted of editors, clinical epidemiologists, and statisticians.[2,4] This standard checklist for reporting of RCTs consists of 21 items regarding the methods, results, and discussion of RCTs. Similar high standards should be required for the report of systematic reviews.[10,20]

It has been acknowledged that reviews of the literature are also susceptible to several types of bias and that a systematic approach may protect against these biases.[26,28] The Cochrane Collaboration is an international network with the objective to prepare, maintain, and disseminate high-quality systematic reviews of RCTs and other sources of evidence of health care interventions.[6] Recently, the Cochrane Back Review Group was established as a subgroup of the Cochrane Musculoskeletal Review Group.[7] It has been shown that many RCTs[17] and reviews[3] in the field of back pain are of low methodologic quality, but also that their reports often lack essential components. In addition, it appears that reviews covering the same topics often reach opposite conclusions.

Because the number of systematic (Cochrane) reviews of interventions for back pain has increased during the past decades and probably will increase more in the near future, it seems useful to develop methodologic guidelines for systematic reviews in this field that address the main steps in conducting a systematic review—namely, a literature search, inclusion criteria, methodologic quality, data extraction, and data analysis. The purpose of these methodologic guidelines is to offer guidance to researchers preparing, conducting, or reporting a systematic review and to readers evaluating these reviews in the field of back pain. They include certain minimum standards the Cochrane Back Review Group expects for reviews to be included in the Cochrane Library. These minimum criteria are items for which some empirical evidence exists that they are associated with bias in systematic reviews,[16,22,29] or for which consensus exists among the editorial board of the Cochrane Back Review Group that they are likely to be associated with bias. In addition, further guidance is presented to enhance the quality of the systematic reviews. The methodologic guidelines not only refer to Cochrane reviews, but are also useful to plan, conduct, or evaluate other systematic reviews (or RCTs) in the field of back pain.

In addition, using these guidelines facilitates comparisons across reviews and enhances consistency among reviewers. It should be emphasized that these guidelines are not a "gold standard" but merely an indication of the current "state-of-the-art" of review methods, which is still in an early stage of evolution. If, in the near future, more empirical evidence will become available on methodologic issues regarding systematic reviews, the Cochrane Back Review Group will update the guidelines presented in this article.

Methods

In December 1996, a survey was sent to the members (n = 9) of the Editorial Board of the Cochrane Back Review Group, ask-

From the *Institute for Research in Extramural Medicine, Faculty of Medicine, Vrije Universiteit, Amsterdam, The Netherlands.
†Co-Editors: C. Bombardier, MD, FRCP, A.L. Nachemson, MD, PhD; Coordinator: R. Esmail, MSc; Other Members: R.A. Deyo, MD, MPH, P.G. Shekelle, MD, MPH, L.M. Bouter, PhD, R.A. de Bie, MSc, G. Waddell, DSc, MD, FRCS, M. Roland, DM, FRCGP, and F. Guillemin, MD, PhD.
Acknowledgment date: July 24, 1997.
Acceptance date: July 30, 1997.

ing them to state which of the items that had been used in Cochrane protocols (n = 15) submitted until December 1996 should be minimally required and which items ought to be considered optional. The items were classified into five categories: literature search, inclusion criteria, methodologic quality, data extraction, and data analysis. In May 1997, a workshop was organized at the Second International Forum for Primary Care Research on Low Back Pain, an invitational conference for approximately 85 established researchers in the field of back pain, in which these methodologic guidelines were discussed. After the workshop, a consensus meeting of the editorial board of the Cochrane Back Review Group took place in which the final version of the guidelines was established. The consensus text was endorsed by detailed comments on drafts of this article in two postal rounds.

■ Method Guidelines

Literature Search

Minimum Criteria. Because one of the main principles of a systematic review is to include all available evidence, the literature search is the first but also a very important step in conducting a systematic review. The search strategy should relate directly to the research question(s) and should be based on the inclusion criteria of the review regarding study design, participants, interventions, outcomes, and language (see Inclusion Criteria section). The Editorial Board of the Cochrane Back Review Group considered carefully which databases should routinely be searched. A simple MEDLINE search is clearly insufficient, because it has been shown that only approximately half of the available RCTs will be identified if the identification of RCTs solely depends on MEDLINE searches. [12] Therefore, we recommend the following as a minimum search strategy. Potentially relevant RCTs meeting the inclusion criteria for a systematic review should be identified by:

1) a computer-aided search of the MEDLINE and EMBASE databases. The highly sensitive search strategies of the U.K. Cochrane Center (Workshop "Advanced Electronic Searching to Identify Reports of Randomized Controlled Trials in MEDLINE and EMBASE," Lefebvre & McDonald, 4th International Cochrane Colloquium, Adelaide, 1996) should be run in conjunction with a specific search for back pain and the intervention at issue (Tables 1 and 2). The MEDLINE search is based on the first two stages of the MEDLINE search strategy recommended by the Cochrane Handbook (Appendix 5C.2) and published by Dickersin et al.[12]
2) a search of the Cochrane Controlled Trials Register that is included in the Cochrane Library (latest issue).
3) communication with the coordinator of the Cochrane Back Review Group (see Conclusion section for address) to identify additional RCTs from the handsearching currently being undertaken by the Cochrane Collaboration and coordinated by the Cochrane Centre in Baltimore.[6]

Table 1. Search Strategy for MEDLINE (OVID)*

Part A: Generic search for randomized controlled trials
randomized controlled trial.pt.
controlled clinical trial.pt.
randomized controlled trials.sh.
random allocation.sh.
double blind method.sh.
single blind method.sh.
1 or 2 or 3 or 4 or 5 or 6
(animal not (human and animal)).sh.
7 not 8
clinical trial.pt.
exp clinical trials/
(clin$ adj25 trial$).ti.ab.
((singl$ or doubl$ or trebl$ or tripl$) adj25 (blind$ or mask$)).ti.ab.
placebos.sh.
placebo$.ti.ab.
random$.ti.ab.
research design.sh.
volunteer$.ti.ab.
10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18
19 not 8
20 not 9
9 or 21
Part B: Specific search for back pain
back pain.sh.
low back pain.sh.
back pain.ti.ab.
backache.ti.ab.
23 or 24 or 25 or 26
Part C: Specific search for intervention at issue

* When using another software package than OVID, consult the manual or your librarian.

4) screening references given in relevant systematic reviews and identified RCTs. Several methods of identification of systematic reviews have been described by Hunt and McKibbon.[15]
5) personal communication with content experts in the field who are requested to review the list of identified RCTs for completeness.

Further Guidance. The first step in the literature search is to decide which articles should be retrieved, ensuring that as many relevant RCTs as possible are identified from the searches in MEDLINE, EMBASE, and possible other databases. We recommend that two reviewers independently apply the inclusion criteria to select the potentially relevant trials from the titles, abstracts, and keywords of the references retrieved by the literature search. We suggest to pilot test the inclusion criteria on a sample of articles, including some considered to be definitely eligible, some definitely not eligible, and some questionable (Cochrane Handbook, Chapter 5.7). [24] A consensus method should be used to solve disagreements between the two reviewers regarding the inclusion of studies. A third reviewer should be consulted if disagreements are not resolved in the consensus meeting. Articles for which disagreement remains and articles for which the title, abstract, and keywords provide insufficient information for a decision on selection should be obtained, to assess whether they meet the inclusion criteria.

If available, a search strategy should preferably also

Table 2. Search Strategy EMBASE (OVID)*

Part A: Generic search for randomized controlled trials
clinical article/
clinical study/
clinical trial/
controlled study/
randomized controlled trial/
major clinical study/
double blind procedure/
multicenter study/
single blind procedure/
phase 3 clinical study/
phase 4 clinical study/
crossover procedure/
placebo/
or/1–13
allocat$.ti.ab.
assign$.ti.ab.
blind$.ti.ab.
(clinic$ adj25 (study or trial)).ti.ab.
compar$.ti.ab.
control$.ti.ab.
cross?over.ti.ab.
factorial$.ti.ab.
follow?up.ti.ab.
placebo$.ti.ab.
prospectiv$.ti.ab.
random$.ti.ab.
((singl$ or doubl$ or trebl$ or tripl$) adj25 (blind$ or mask$)).ti.ab.
trial.ti.ab.
(versus or vs).ti.ab.
or/15–29
14 or 30
human/
nonhuman/
animal/
animal experiment/
33 or 34 or 35
32 and 36
31 not 36
31 and 37
38 or 39
Part B: Specific search for back pain
backache/
low back pain/
back pain.ti.ab.
backache.ti.ab.
or/41–44
Part C: Specific search for intervention at issue

* When using another software package than OVID, consult the manual or your librarian.

be run in a specific database for the intervention at issue (for example, CHIROLARS[1] or the database of the Cochrane Field of Rehabilitation and Therapy.[25]

Citation tracking of the identified RCTs (use of Science Citation Index to search forward in time which subsequent articles have cited the identified RCTs) should be considered. The value of using citation tracking has not yet been established, but it may be especially useful to identify additional studies on topics that are poorly indexed in MEDLINE and EMBASE.

Inclusion Criteria

Minimum Criteria. *Study Design*—RCTs should be included.

Participants—Participants of RCTs that will be included in the systematic review should be defined explicitly in terms of age, gender, duration of symptoms, localization of symptoms, and type of symptoms.

Example of definition of participants:

> Randomized controlled trials were included that reported on subjects with nonspecific low back pain and between 18 and 65 years of age. Low back pain was defined as pain localized below the lowest ribs and above the inferior gluteal folds; nonspecific indicates that no specific cause is detectable, such as infection, neoplasm, metastasis, osteoporosis, rheumatoid arthritis, fracture, inflammatory process, or radicular syndrome. Randomized controlled trials that include subjects with specific low back pain caused by specific pathologic entities such as infection, neoplasm, metastasis, osteoporosis, rheumatoid arthritis, fracture, inflammatory process, or radicular syndrome were excluded.

Interventions—The type of interventions that will be included in the systematic review should be explicitly described.

Example of description of interventions:

> Randomized controlled trials in which one of the treatments consisted of a back school type of intervention were included. A back school is defined as consisting of an education and skills program, including exercises, in which all lessons are given to groups of patients and supervised by a paramedical therapist or medical specialist. Additional interventions are allowed.

Outcomes—The outcome measures and instruments that will be included in a systematic review should be explicitly described. Preferably, the outcome measures should be simple, relevant, practical, symptom-specific, and widely applicable, and the instruments should be reliable, valid, and responsive. The Editorial Board of the Cochrane Back Review Group considered that outcomes of symptoms (*e.g.*, pain), overall improvement or satisfaction with treatment, function (*e.g.*, back specific functional status), well-being (*e.g.*, quality of life) and disability (*e.g.*, absenteeism or return to work) should be included in a systematic review of back pain if they are reported in the original RCTs.

Example of definition of outcome measures and instruments:

> Randomized controlled trials were included that use at least one of the four outcome measures we consider to be the most important—pain (Visual Analogue Scale, numerical rating scale), a global measure (overall improvement, proportion of patients recovered, subjective improvement of symptoms), back specific functional status (*e.g.*, Roland Disability Questionnaire, Oswestry Scale), and return to work (return to work status, days off work).

Further Guidance. *Study Design*—Nonrandomized controlled clinical trials might be included if the available evidence from RCTs is not sufficient (*e.g.*, because there are very few RCTs or the RCTs are of poor methodologic

quality). If authors wish to extend a systematic review beyond RCTs, the reason for this should be outlined.

Outcomes—Outcomes of physical examination (range of motion, spinal flexibility, degrees of straight leg raising, or muscle strength), other symptoms (medication use, health care use, side effects) and economic outcomes may be included where appropriate, considering the aim of the intervention at issue.

Language—Inclusion of studies published in other languages than English is recommended, although we acknowledge that it may not always be feasible and may depend on the time and resources available. However, there is some empirical evidence that exclusion of trials published in other languages than English might be associated with bias. Grégoire et al.[14] assumed that positive results by authors from non-English speaking countries are more likely to be published in English and negative results in the authors' language. They found an example of a meta-analysis where inclusion of a non-English language trial changed the results and conclusion.[14] Moher et al.[21] evaluated the quality of reporting of RCTs published in English, French, German, Italian, and Spanish between 1989 and 1993 and did not find significant differences. The authors of both publications[14,21] concluded that all trials should be included in a systematic review regardless of the language in which they were published, to increase precision and reduce bias. If RCTs published in other languages are excluded from the review, the reason for this decision should be given. Especially on topics where there are likely to be a significant number of non-English language publications (for example, the Asian literature on acupuncture), it may be wise to consider involvement of a collaborator with relevant language skills.

Methodologic Quality

Minimum Criteria. The methodologic quality of the studies should be independently assessed by at least two reviewers.

Currently, there is still limited empirical evidence of a relation between specific methodologic criteria and bias. Some authors have reported that inadequate concealment of treatment allocation is associated with larger effect sizes,[8,29] whereas others reported a bias in the opposite direction.[9] Some empirical evidence has been found for a relation between double-blinding and bias.[9,29] The Editorial Board of the Cochrane Back Review Group considered that the methodologic criteria list should minimally identify whether the following items were features of each trial: treatment allocation, withdrawal/drop-out rate, patients blinded, outcome assessor blinded, and intention-to-treat analysis. The operationalization of the criteria should be described explicitly (see, for example, Table 4), and the criteria should be scored as positive, negative, or unclear ("yes," "no," and "don't know").

Further Guidance. Some empirical evidence has been provided that blind assessment of the methodologic quality of trials (where studies were blinded with regard to authors, institution, and journal) resulted in lower and more consistent scores than open assessment.[16] However, in a recently conducted study, researchers did not find an association between the assessment of studies blinded for authors, institution, journal, results, and conclusions on one side and bias on the other.[32] Although it is difficult to achieve true blinding, because experts are usually involved in the assessment of the methodologic quality of the studies, we recommend that the author, institution, and journal be removed from copies when assessing the methodologic quality. Because the quality assessment of content experts may be biased by prior opinions, it may be desirable to have both an expert and a non-expert (but with a research background) assess the quality of the studies. This is also one of the reasons we consider independent assessment by at least two reviewers as a minimum criterion. In systematic reviews where there is likely to be a conflict of interest (*e.g.*, pharmaceutical companies reviewing drugs, chiropractors reviewing spinal manipulation, or physiotherapists reviewing exercise therapy), it may be desirable to blind the studies also for results and conclusions or to include someone who has no conflict of interest in the assessment of the quality of the studies.

We recommend that reviewers pilot test the methodologic quality assessment on some similar articles (regarding another intervention or disorder) that will not be included in the review, including articles of presumed low, moderate, and high quality. Experience has shown that even researchers skilled in systematic reviews make systematically different interpretations of the quality assessment criteria, and it is important for reviewers to agree on a common interpretation.

We recommend using a consensus method to discuss and solve the disagreements between the reviewers. If disagreement persists, an additional independent person should be consulted. The interobserver reliability (*e.g.* Kappa) of the quality assessment might be evaluated and reported.

If the article does not contain information on (one or more of) the methodologic criteria, the authors might be contacted for additional information. If the authors cannot be contacted or if the information is no longer available, the criteria should be scored as "unclear." If many studies are scored as "unclear" for the same criterion, the reviewers could consider dropping that criterion *post hoc* (Chapter 6 of the Cochrane Handbook).[24]

We recommend using a specific criteria list (Table 3) and a uniform operationalization of criteria (Table 4) in systematic reviews of back pain. We recognize that a number of these criteria require judgment by the reviewer, and this is one of the reasons we recommend independent assessment by at least two reviewers as a minimum criterion (see above). Currently, there is s

Table 3. Criteria List for the Methodologic Quality Assessment

Patient selection		
a.	Were the eligibility criteria specified?	Yes/No/Don't know
b.	Treatment allocation	
	1) Was a method of randomization performed?	Yes/No/Don't know
	2) Was the treatment allocation concealed?	Yes/No/Don't know
c.	Were the groups similar at baseline regarding the most important prognostic indicators?	Yes/No/Don't know
Interventions		
d.	Were the index and control interventions explicitly described?	Yes/No/Don't know
e.	Was the care provider blinded to the intervention?	Yes/No/Don't know
f.	Were co-interventions avoided or comparable?	Yes/No/Don't know
g.	Was the compliance acceptable in all groups?	Yes/No/Don't know
h.	Was the patient blinded to the intervention?	Yes/No/Don't know
Outcome measurement		
i.	Was the outcome assessor blinded to the intervention?	Yes/No/Don't know
j.	Were the outcome measures relevant?	Yes/No/Don't know
k.	Were adverse effects described?	Yes/No/Don't know
l.	Was the withdrawal/drop-out rate described and acceptable?	Yes/No/Don't know
m.	Timing follow-up measurements	
	1) Was a short-term follow-up measurement performed?	Yes/No/Don't know
	2) Was a long-term follow-up measurement performed?	Yes/No/Don't know
n.	Was the timing of the outcome assessment in both groups comparable?	Yes/No/Don't know
Statistics		
o.	Was the sample size for each group described?	Yes/No/Don't know
p.	Did the analysis include an intention-to-treat analysis?	Yes/No/Don't know
q.	Were point estimates and measures of variability presented for the primary outcome measures?	Yes/No/Don't know

Internal validity criteria: b, e, f, g, h, i, j, l, n, p.
Descriptive criteria: a, c, d, k, m.
Statistical criteria: o, q.

no empirical evidence of an association with bias for most methodologic criteria except adequate concealment of treatment allocation and double-blinding.[8,9,29] In addition, this evidence is collected in other fields than back pain and it is still unclear whether this evidence is also valid for back pain studies.

Moher et al.[19] identified more than 30 criteria lists that have been used to assess the methodologic quality of RCTs, with the number of criteria ranging from 3 to 35. Most of these lists have been developed without a well-defined conceptual model and have not been thoroughly investigated with regard to their reliability and validity. Recently, two instruments for methodologic quality assessment in RCTs have been developed according to established methodologic procedures, one which resulted in a final list of three criteria[16] and the other of nine criteria.[31] We recommend using the list presented in Table 3, which is a modified version of a list that has already been used in a number of systematic reviews in the field of back pain[30] and physiotherapy[33] and includes all criteria of the lists of Jadad et al.[16] and Verhagen et al.[31] The list consists of internal validity criteria, descriptive criteria, and statistical criteria. The internal validity criteria (n = 10) refer to characteristics of the study that might be related to selection bias (criterion b), performance bias (criteria e, f, g, and h), attrition bias (criteria l and p), and detection bias (criteria i, j, and n), and should be used to define methodologic quality in the meta-analysis. The descriptive criteria (criteria a, c, d, k, and m) refer to the external validity of the study and may be used for the subgroup and sensitivity analyses (see Data Analysis section). The statistical criteria (criteria o and q) indicate whether calculations can be made and conclusions can be drawn independently of the opinion of the authors of the original study.

Differences in methodologic quality may explain variation in the results of the studies included in a systematic review. Quantitative (statistical pooling) or qualitative meta-analysis of studies with different methodologic quality can result in over- or underestimation of the effectiveness of the intervention at issue. However, there are no strict guidelines for the use of methodologic quality assessment in systematic reviews. In general, we recommend choosing from the options listed below[11,18,24] and clearly describing the reasoning behind the choice. First, the methodologic quality can be used as an additional criterion for inclusion of studies in the review based on one or more items (*e.g.*, only inclusion of RCTs or double-blind RCTs) or based on a summary quality score (*e.g.*, only studies that adequately fulfill 50% or more of the validity criteria). Choosing cut-off points for inclusion or exclusion of studies remains arbitrary. Second, a stratified analysis can be performed in which the results are separately presented for different strata of studies (*e.g.*, low and high methodologic quality studies defined by a preset cut-off point). Third, a sensitivity analysis can be performed to determine whether the overall results are the same when studies above different methodologic cut-off points are analyzed. Fourth, a cumulative meta-analysis can be performed by examining the impact on the overall results as studies of decreasing quality are included one at a time. Fifth, a meta-regression can be performed to explore the relation between methodologic quality and the magnitude of effect across studies.[5,13] And sixth, weights can be applied in the analysis to studies according to the methodologic quality, so that studies of higher quality have more impact on the overall results. Obviously, choosing weights also involves some arbitrary judgment.

Data Extraction

Minimum Criteria. At least two reviewers should independently extract the data. The data that should be extracted include characteristics of participants (*e.g.*, age, gender, type and duration of disorder), interventions (type, dose or intensity, frequency, and duration) and outcomes (type of outcome measure and instrument) and results. For dichotomous data, the number of people who experienced the outcome in each group and the total number in each group should be extracted. For continu-

Table 4. Operationalization of the Criteria List

a. To score a "yes," the radiation pattern of back pain and the duration of the disorder must be described appropriately.
b1. A random (unpredictable) assignment sequence. Methods of allocation using date of birth, date of admission, hospital numbers, or alternation should not be regarded as appropriate.
b2. Assignment generated by an independent person not responsible for determining the eligibility of the patients. This person has no information about the persons included in the trial and has no influence on the assignment sequence or on the decision about eligibility of the patient.
c. To receive a "yes," groups must be similar at baseline regarding age, duration of complaints, percentage of patients with radiating pain, and value of main outcome measure(s).
d. Adequate description of type, modality, application technique, intensity, duration, number of frequency of sessions for both the index intervention and control intervention(s), so that others could replicate the treatment.
e. The reviewer determines when enough information about the blinding is given in order to score a "yes."
f. Co-interventions should either be avoided in the trial design or comparable between the index and control groups.
g. The reviewer determines when the compliance to the interventions is acceptable, based on the reported intensity, duration, number, and frequency of sessions for both the index intervention and control intervention(s).
h. The reviewer determines when enough information about the blinding is given in order to score a "yes."
i. The reviewer determines (per outcome parameter) when enough information about blinding is given in order to score a "yes."
j. The reviewer determines whether the outcome measures were relevant. For back pain, we recommend considering pain, a global measure of improvement, back specific functional status, generic functional status, and return-to-work to be relevant.
k. Each event should be described and correctly attributed to the allocated treatment: if it is explicitly reported that "no adverse effects" have occurred, a "yes" should be scored.
l. Participants included in the study but who did not complete the observation period or were not included in the analysis must be described. If the percentage of withdrawals and drop-outs does not exceed 20% for short-term follow-up and 30% for long-term follow-up and does not lead to substantial bias, a "yes" is scored. (N.B., these percentages are arbitrary, not supported by literature).
m1. Outcome assessment at the end of the intervention period.
m2. Outcome assessment >3 months after randomization.
n. Timing of outcome assessment should be identical for all intervention groups and for all important outcome assessments.
o. To be presented for each group at randomization and for most important outcome assessments; N.B., this means that, in contrast to previous lists, there is no pre-set cut-off point to determine whether the sample size is sufficient.
p. All randomized patients are reported/analyzed for the most important moments of effect measurement (minus missing values) irrespective of non-compliance and co-interventions.
q. Both point estimates and measures of variability should be presented (to be scored for each important outcome parameter separately). Point estimates are: means, medians, modes, etc.: Measures of variability are: standard deviations, 95% confidence intervals, etc.

ous data, the number of people, the mean value, and the standard deviation (or the median value and interquartile range if the outcome is not normally distributed) for the outcome in each group should be extracted. A distinction should be made between data collected but not reported and data not collected by the original authors.

Further Guidance. Similar to the assessment of the methodologic quality, we recommend that the data extraction be performed with the studies blinded for authors, journal, and institution until further evidence becomes available. We recommend using a standardized form for data extraction that will facilitate the comparison process. It is advisable to pilot test the data extraction form to avoid any misinterpretations or later disagreements. If there are any disagreements, consensus should be achieved by discussion among the reviewers. If disagreements persist, an additional independent person should be consulted.

Data extraction forms will vary across different systematic reviews, but there will also be similarities among the forms needed for reviews on back pain. Because designing a data extraction form is time-consuming, and given the important function of data extraction forms, it might be helpful to profit and learn from experiences of others. Copies of data extraction forms of back pain reviews included in the Cochrane Library can be obtained from the coordinator of the Cochrane Back Review Group, on request.

Data Analysis

Minimum Criteria. As in every scientific study, statistical methods can also be used to analyze and summarize the data in a systematic review. The objective of most systematic reviews will be to provide a reliable overall estimate of the effects of an intervention, based on a weighted average of the results of all the available studies. There is consensus among the Editorial Board of the Cochrane Back Review Group that if relevant, valid data are lacking (data are too sparse or of too low quality) or if data are statistically and clinically too heterogeneous, a meta-analysis should be avoided and reviewers should perform a qualitative review. In these instances, a qualitative review might be performed by attributing various levels of evidence to the effectiveness of a treatment, taking into account the participants, interventions, outcomes, and methodologic quality of the original studies.[30] Because various approaches are possible in analyzing the data in a systematic review, we suggest that authors clearly outline the reasoning behind the approach they use.

Further Guidance. The Editorial Board of the Cochrane Back Review Group suggests referring to the recommendations of Chapter 8 of the Cochrane Handbook for further guidance on data analysis.[24]

The comparisons with regard to the outcomes and subgroups that will be analyzed should relate directly to the objective or research question(s) of the review. If it is clinically relevant and statistically justified to combine

the results, a meta-analysis should be performed that provides an overall effect estimate, with a 95% confidence interval for each comparison.[34] For the meta-analysis of dichotomous outcomes, the relative risk, risk difference, or odds ratio can be used to summarize the effect. For continuous outcomes, mean differences from each trial can be combined. If the continuous outcomes are not directly combinable—that is, if different instruments are used for the outcome measurements—standardized mean differences (effect sizes) might be used.[34] If necessary, the authors of the original studies should be contacted to provide relevant information.

There are two statistical models for combining data in a meta-analysis: the fixed effects model and the random effects model.[13,23] Although there are arguments favoring each model, in general, the clinical heterogeneity of the back pain literature suggests that the assumptions underlying the random effects model are better suited to statistical combinations of different RCTs in this field. An explicit list should be given of the characteristics of participants in, the types of interventions for, and the exact outcome values from, each study of a group of studies that are combined. Sensitivity analyses should be performed to examine the impact of variation in validity scores of the different groups of studies (see Methodologic Quality section).

It may be difficult sometimes for reviewers to decide whether it is clinically relevant to combine the results of a group of studies in a meta-analysis—for example, studies of patients with different types of treatments, different types of comparison groups, or different clinical characteristics of patients studied. There are no simple answers here, and reviewers will need to be explicit about their decisions so that others may judge for themselves whether these choices were clinically sensible. A related but separate issue concerns statistical homogeneity. A test for the statistical homogeneity of studies may be performed to evaluate whether the differences among the results of the studies are greater than by chance alone (Cochrane Handbook, Chapter 8.3).[24] However, the test is not powerful, and failure to reject the hypothesis of homogeneity is not proof that the studies are homogeneous. If the hypothesis of homogeneity is rejected, or if the reviewers decide, on clinical grounds, that the studies are too heterogeneous to support statistical combinations, then the potential sources of heterogeneity should be examined because the observed differences might be caused by factors other than chance, such as differences in methodologic quality, characteristics of participants, interventions, or outcomes.[10] If the heterogeneity can be explained, reviewers should consider presenting the results of the relevant subgroups at issue. The subgroup analyses should be kept to a minimum and should be defined *a priori*, because subgroup analyses can be informative but also misleading.[27,35] If the heterogeneity cannot be explained, reviewers should perform a qualitative analysis (see minimum criteria).

■ Discussion

Currently, there seems to be a general acceptance of the need for, and importance of, systematic reviews. The Editorial Board of the Cochrane Back Review Group believes that systematic reviews represent one of the key advances in medical science in the past 10 years and offer the real opportunity to lead to changes in medical practice worldwide. Systematic reviews need to be conducted as carefully as the trials they report and, to achieve full impact, systematic reviews need to meet high methodologic standards. The objective of these methodologic guidelines is to help reviewers design, conduct, and report reviews of trials in the field of back pain systematically and explicitly. These guidelines are not intended to dictate a gold standard or to discourage people from doing a systematic review. On the contrary, we would like to encourage people to undertake a systematic review in collaboration with others. For more guidance on systematic reviews of back pain, we recommend referring to the Cochrane Handbook or contacting the Cochrane Back Review Group coordinator, Rosmin Esmail, Institute for Work & Health, Toronto, Ontario, Canada, telephone: (416) 927-2027, fax: (416) 927-4167, e-mail: resmail@iwh.on.ca.

References

1. Aker PD, McDermaid C, Opitz BG, White MW. Searching chiropractic literature: A comparison of three computerized databases. J Manipulative Physiol Ther 1996;19:518–24.
2. Altman DG. Better reporting of randomised controlled trials: The CONSORT statement. Br Med J 1996;313:570–1.
3. Assendelft WJJ, Koes BW, Knipschild PG, Bouter LM. The relationship between methodological quality and conclusions in reviews of spinal manipulation. JAMA 1995;274:1942–8.
4. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: The CONSORT statement. JAMA 1996;276:637–9.
5. Berlin JA, Antman EM. Advantages and limitations of meta-analytic regressions of clinical trials data. Online J Curr Clin Trials 1994;Doc. 134.
6. Bero L, Rennie D. The Cochrane Collaboration: Preparing, maintaining, and disseminating systematic reviews of the effects of health care. JAMA 1995;274:1935–8.
7. Bombardier C, Esmail R, Nachemson AL. The Cochrane Collaboration Back Review Group for Spinal Disorders. Spine 1997;22:837–40.
8. Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. N Engl J Med 1983; 309:1358–61.
9. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. Stat Med 1989;8:441–54.
10. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam consultation on meta-analysis. J Clin Epidemiol 1995;48:167–71.
11. Detsky AS, Naylor CD, O'Rourke K, McGreer AJ, L'Abbé KA. Incorporating variations in the quality of individual ran-

domized trials into meta-analysis. J Clin Epidemiol 1992;45: 255–65.
12. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. Br Med J 1994;309:1286–91.
13. Greenland S. Invited commentary: A critical look at some popular meta-analytic methods. Am J Epidemiol 1994;140: 290–6.
14. Grégoire G, Derderian F, Le Lorier J. Selecting the language of the publications included in a meta-analysis: Is there a tower of Babel bias? J Clin Epidemiol 1995;48:159–63.
15. Hunt DL, McKibbon KA. Locating and appraising systematic reviews. Ann Intern Med 1997;126:532–8.
16. Jadad AR, Moore A, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clin Trials 1996;17:1–12.
17. Koes BW, Bouter LM, van der Heijden GJMG. Methodological quality of randomized clinical trials on treatment efficacy in low back pain. Spine 1995;20:228–35.
18. L'Abbé KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. Ann Intern Med 1987;107:224–33.
19. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Control Clin Trials 1995;16:62–73.
20. Moher D, Olkin I. Meta-analysis of randomized controlled trials: A concern for standards. JAMA 1995;274:1962–4.
21. Moher D, Fortin P, Jadad AR, et al. Completeness of reporting of trials published in languages other than English: Implications for conduct and reporting of systematic reviews. Lancet 1996;347:363–6.
22. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: Current issues and future directions. Int J Technol Assessment Health Care 1996;12:195–208.
23. Morris RD. Meta-analysis in cancer epidemiology. Environ Health Perspect 1994;102 (suppl 8):61–6.
24. Mulrow CD, Oxman AD (eds). Cochrane Collaboration Handbook [updated 1 March 1997]. In: The Cochrane Library [database on disk and CDROM]. The Cochrane Collaboration. Oxford: Update Software, 1996.
25. Newham D. The Cochrane Collaboration: What is it, how does it work, and what has it to do with physiotherapy. Physiotherapy 1995;81:405–7.
26. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol 1991;44:1271–8.
27. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992;116:78–84.
28. Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. JAMA 1994; 272:1367–71.
29. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408–12.
30. Van Tulder MW, Koes BW, Bouter LM (eds). Low back pain in primary care: Effectiveness of diagnostic and therapeutic interventions. EMGO-Institute, Vrije Universiteit, Amsterdam, The Netherlands, 1996.
31. Verhagen AP, de Vet HCW, de Bie RA, et al. The Delphi list: A criteria list for quality assessment of randomized clinical trials developed by Delphi consensus. (Submitted).
32. Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Knipschild PG. Balneotherapy and quality assessment: The interobserver reliability of the Maastricht criteria list and the need of blinded quality assessment. (Submitted).
33. Vet HCW de, de Bie RA, van der Heijden GJMG, Verhagen AP, Sijpkes P, Knipschild PG. Systematic reviews on the basis of methodological criteria. Physiotherapy 1997;83: 284–9.
34. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. Stat Med 1991;10:1665–77.
35. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991;266:93–8.

Address reprint requests to

Maurits W. van Tulder, PhD
Institute for Research in Extramural Medicine
Faculty of Medicine
Vrije Universiteit, van der Boechorststraat 7
1081 BT, Amsterdam, The Netherlands
E-mail: mw.van_tulder.emgo@med.vu.nl